A Simple Trick for Constructing Bayesian Formulations of Sparse Kernel Learning Methods

Gavin C. Cawley
School of Computing Sciences
University of East Anglia
Norwich NR4 7TJ U.K.
E-mail: gcc@cmp.uea.ac.uk

Nicola L. C. Talbot
School of Computing Sciences
University of East Anglia
Norwich NR4 7TJ U.K.
E-mail: nlct@cmp.uea.ac.uk

***Abstract*— In this paper, we present a simple mathematical trick that simplifies the derivation of Bayesian treatments of a variety of sparse kernel learning methods. The incomplete Cholesky factorisation due to [1] is used to transform the dual parameter space, such that the covariance matrix of the Gaussian prior over model parameters becomes the identity matrix. The regularisation term is then the familiar weight-decay regulariser, allowing the Bayesian analysis to proceed straight-forwardly via the methods developed by [2–4]. As a bye-product, the *incomplete* Cholesky factorisation algorithm also identifies a subset of the training data forming an approximate basis for the remaining data in feature space, resulting in a sparse model. Bayesian treatments of the kernel ridge regression algorithm [5], with both constant and input dependent variance structures, are given as illustrative examples of the proposed technique, which we hope will be more widely applicable.**

I. Introduction

The "kernel trick" provides a mathematically elegant means of constructing powerful non-linear variants of classical (linear) statistical methods, such as ridge regression [5, 6], Fisher discriminant analysis [7, 8] and principal component analysis [9, 10], as well as forming an integral component of more recent developments such as the support vector machine [11, 12]. As a consequence of the linear nature of the underlying learning algorithm, the optimal values of the primary model parameters can normally be found very efficiently. However, the generalisation performance of kernel learning methods is often heavily dependant on the values of kernel and regularisation parameters, which are most often determined via minimisation of the cross-validation error [13] or theoretical bounds on test error (e.g. [14]). Bayesian model selection procedures have also been proposed for the support vector machine [15] and least-squares support vector machine [16], based on the evidence framework of [2–4]. This approach is attractive as it provides a theoretically sound means of selecting good values for all hyper-parameters, based solely on the training data. The Bayesian approach can also be used to provide a credible interval on model predictions.

In this paper, we present a simple trick allowing the evidence framework to be applied to a broad class of sparse kernel learning methods. For sparse kernel machines, the non-spherical covariance structure of the Gaussian prior over model parameters is inconvenient in estimating the number of effective parameters used in updating the regularisation parameters. Therefore we simply transform the parameter space such that the prior becomes spherical, the evidence framework can then be applied without further modification. The required transformation and simultaneous sparsification of the kernel machine are achieved using the incomplete Cholesky factorisation algorithm due to [1]. Sparsity is an important issue in large-scale applications of kernel learning methods as the computational complexity of the training procedures are often as high as $\mathcal{O}(\ell^3)$ operations, where ℓ is the number of training patterns, whereas the sparse Bayesian methods considered here typically have computational complexities of only $\mathcal{O}(\ell W^2)$, where $W \ll \ell$ is the number of non-zero parameters.

II. Method

A wide range of kernel learning methods, given training data

$$\mathcal{D} = \{(\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d, \quad t_i \in \mathcal{T} \subset \mathbb{R})\}_{i=1}^{\ell},$$

construct a linear model[1] $y(\boldsymbol{x};\boldsymbol{w}) = f\{\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x})\}$ in a fixed feature space, $\mathcal{F}$ $(\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F})$, that captures some statistical dependency between pairs of patterns in $\mathcal{X}$ and $\mathcal{T}$. Rather than specify the feature space explicitly, the feature space is induced by a kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defining the inner product between vectors in $\mathcal{F}$, i.e. $\mathcal{K}(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})\cdot\boldsymbol{\phi}(\boldsymbol{x}')$. A commonly encountered kernel function is the Radial Basis Function (RBF) kernel,

$$\mathcal{K}(\boldsymbol{x},\boldsymbol{x}') = \exp\left\{\sum_{i=1}^{d}\eta_i(x_i - x_i')^2\right\}, \tag{1}$$

where the *kernel parameters*, $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)$, provide individual control over the sensitivity of the kernel function to each input variable. The optimal values of the model parameters $\boldsymbol{w}$ are determined by minimising a regularised loss function [17],

$$S(\boldsymbol{w}) = \sum_{i=1}^{\ell} C\left\{y(\boldsymbol{x}_i;\boldsymbol{w}), t_i\right\} + \zeta\frac{1}{2}\|\boldsymbol{w}\|^2, \tag{2}$$

[1]The usual bias parameter, b, is omitted here purely to improve the clarity of later derivations, but is included in practical implementations described in section III.

where $C\{\cdot,\cdot\}$ is a convex function measuring the misfit between the output of the model and the desired output, for example the sum-of-squares error $C\{y(\boldsymbol{x};\boldsymbol{w}),t\} = [t - y(\boldsymbol{x};\boldsymbol{w})]^2$. The *regularisation parameter*, ζ, controls the bias-variance trade-off [18], and must be carefully adjusted in order to avoid over-fitting the training data. The representer theorem [19] indicates that the solution to an optimisation problem of this nature can be expressed in the form

$$y(\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_i,\boldsymbol{x}). \tag{3}$$

The advantage of the "kernel trick" is then apparent; a linear model can be constructed in a high- or even infinite-dimensional feature space, resulting in a very flexible non-linear model, whilst involving only finite-dimensional quantities, principally the vector of model parameters $\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\ldots,\alpha_\ell)$, and the *Gram* matrix $\boldsymbol{K} = \{k_{ij} = \mathcal{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)\}_{i,j=1}^{\ell}$. The regularisation term can then be re-written in terms of $\boldsymbol{\alpha}$ and $\boldsymbol{K}$ as follows,

$$\boldsymbol{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i) \implies \|\boldsymbol{w}\|^2 = \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}. \tag{4}$$

A. Bayesian Interpretation

Let $E_{\mathcal{D}}$ and $E_{\mathcal{W}}$ represent the contributions to the cost function (2) due to the data misfit and regularisation terms respectively,

$$E_{\mathcal{D}} = \sum_{i=1}^{\ell} C\left\{y(\boldsymbol{x}_i;\boldsymbol{\alpha}),t_i\right\} \quad \text{and} \quad E_{\mathcal{W}} = \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}.$$

Minimisation of the cost function (2) is then equivalent to maximising the Bayesian posterior distribution over model parameters,

$$p(\boldsymbol{\alpha}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{p(\mathcal{D})},$$

where the *likelihood* of the data with respect to $f(\cdot;\boldsymbol{\alpha})$ is given by

$$p(\mathcal{D}|\boldsymbol{\alpha}) = \frac{1}{Z_{\mathcal{D}}(\boldsymbol{\alpha})} \exp\left\{-E_{\mathcal{D}}\right\}$$

and the prior distribution over model parameters by

$$p(\boldsymbol{\alpha}) = \frac{1}{Z_{\mathcal{W}}(\boldsymbol{\alpha})} \exp\left\{-\zeta E_{\mathcal{W}}\right\},$$

where $Z_{\mathcal{D}}(\boldsymbol{\alpha})$ and $Z_{\mathcal{W}}(\boldsymbol{\alpha})$ represent appropriate normalising functions. For kernel learning methods, the prior distribution over model parameters is then a zero-mean multi-variate Gaussian distribution with covariance matrix $\boldsymbol{K}^{-1}$,

$$p(\boldsymbol{\alpha}) = \frac{1}{\sqrt{(2\pi/\zeta)^\ell |\boldsymbol{K}|}} \exp\left\{-\frac{1}{2}\zeta\boldsymbol{\alpha}\boldsymbol{K}\boldsymbol{\alpha}\right\}. \tag{5}$$

B. Sparse Bayesian Kernel Ridge Regression

Kernel Ridge regression [5] implements a form of regularised non-linear least-squares regression, where the data misfit term in the optimisation criterion (2) is given by

$$E_{\mathcal{D}} = \frac{1}{2}\sum_{i=1}^{\ell}\left[t_i - y(\boldsymbol{x}_i,\boldsymbol{\alpha})\right]^2.$$

For large scale applications, optimising the vector of coefficients of a fully dense kernel expansion (3) may not be computationally feasible. Instead we may opt to approximate (3) by a sparse expansion where only a fraction of the coefficients assume non-zero values. Without loss of generality, we will for the moment assume that the first W coefficients are retained, and return to the issue of choosing a good set of coefficients at a later stage. The output of the sparse kernel machine is then given by

$$\tilde{y}(\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{i=1}^{W} \alpha_i \mathcal{K}(\boldsymbol{x}_i,\boldsymbol{x}).$$

The optimisation criterion to be minimised in training a sparse kernel ridge regression model then becomes,

$$S(\boldsymbol{\alpha}) = \frac{\xi}{2}\sum_{i=1}^{\ell}\left[t_i - \tilde{y}(\boldsymbol{x}_i;\boldsymbol{\alpha})\right]^2 + \frac{\zeta}{2}\sum_{j=1}^{W}\alpha_j^2.$$

The most probable vector of model parameters, $\boldsymbol{\alpha}_{\text{MP}}$, i.e. those minimising $S(\boldsymbol{\alpha})$ or equivalently maximising the posterior distribution, are given by the solution of a system of linear equations,

$$\left[\xi\hat{\boldsymbol{K}}^T\hat{\boldsymbol{K}} + \zeta\tilde{\boldsymbol{K}}\right]\boldsymbol{\alpha} = \xi\hat{\boldsymbol{K}}^T\boldsymbol{t}, \tag{6}$$

where $\tilde{\boldsymbol{K}} = [\mathcal{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{i,j=1}^{W}$ is the symmetric sub-matrix of $\boldsymbol{K}$ formed by the first W columns of the first W rows, and $\hat{\boldsymbol{K}} = [\mathcal{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{i,j=1}^{i=W,j=\ell}$ is the sub-matrix comprised of the first W columns of $\boldsymbol{K}$. The likelihood of the i.i.d. training data with respect to the model parameters, $\boldsymbol{\alpha}$, is a Gaussian distribution,

$$p(\mathcal{D}|\boldsymbol{\alpha}) = \frac{1}{Z_{\mathcal{D}}(\xi)}\exp\left\{-\frac{\xi}{2}\sum_{i=1}^{\ell}\left[t_i - \tilde{y}(\boldsymbol{x}_i;\boldsymbol{\alpha})\right]^2\right\}$$

where ξ represents the inverse variance of the assumed additive zero-mean Gaussian noise process, i.e.

$$t_i = y(\boldsymbol{x}_i;\boldsymbol{\alpha}) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,\xi^{-1}).$$

The posterior distribution over the model parameters, $\boldsymbol{\alpha}$, is also a Gaussian distribution, centred on the most probable vector of coefficients, $\boldsymbol{\alpha}_{\text{MP}}$,

$$p(\boldsymbol{\alpha}|\mathcal{D}) = \frac{1}{Z_{\mathcal{S}}}\exp\left\{-S(\boldsymbol{\alpha}_{\text{MP}}) - \frac{1}{2}\Delta\boldsymbol{\alpha}^T\boldsymbol{A}\Delta\boldsymbol{\alpha}\right\}$$

where $\Delta\boldsymbol{\alpha} = \boldsymbol{\alpha} - \boldsymbol{\alpha}_{\text{MP}}$ and $\boldsymbol{A}$ is the Hessian of $S(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, evaluated at the most probable value,

$$\boldsymbol{A} = \nabla\nabla S(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}_{\text{MP}}} = \xi\hat{\boldsymbol{K}}^T\hat{\boldsymbol{K}} + \zeta\tilde{\boldsymbol{K}}. \tag{7}$$

C. The Evidence Framework for ξ and ζ

The evidence approximation of [2–4] assumes that the posterior distribution for the hyper-parameters, $p(\zeta, \xi|\mathcal{D})$, is sharply peaked about their most probable values, $\zeta_{\rm MP}$ and $\xi_{\rm MP}$, suggesting the following approximation to the posterior distribution for $\boldsymbol{\alpha}$,

$$\begin{aligned} p(\boldsymbol{\alpha}|\mathcal{D}) &= \iint p(\boldsymbol{\alpha}|\zeta, \xi, \mathcal{D}) p(\zeta, \xi|\mathcal{D}) d\zeta d\xi \\ &\approx p(\boldsymbol{\alpha}|\zeta_{\rm MP}, \xi_{\rm MP}, \mathcal{D}). \end{aligned}$$

Thus, rather than integrate out the regularisation parameter entirely (e.g. [20]), we simply proceed with the analysis using the regularisation parameter fixed at its most likely value. For a discussion of the validity of this approach, see MacKay [21]. We seek therefore to maximise the posterior distribution,

$$p(\zeta, \xi|\mathcal{D}) = \frac{p(\mathcal{D}|\zeta, \xi) p(\zeta, \xi)}{p(\mathcal{D})}.$$

If the prior, $p(\zeta, \xi)$ is relatively insensitive to the values of ζ and ξ, then maximising the posterior is approximately equivalent to maximising the likelihood term, $p(\mathcal{D}|\zeta, \xi)$, known as the *evidence* for ζ and ξ. Assuming the the posterior for the model parameters is Gaussian, the log-evidence is given by

$$\begin{aligned} \log p(\mathcal{D}|\zeta, \xi) &= \frac{W}{2} \log \zeta + \frac{\ell}{2} \log \xi - \frac{\ell}{2} \log(2\pi) \\ &- \xi E_{\mathcal{D}}^{\rm MP} - \zeta E_{\mathcal{W}}^{\rm MP} - \frac{1}{2} \log |\boldsymbol{A}|. \end{aligned} \quad (8)$$

Update formula for each of the hyper-parameters are normally then obtained by setting the derivative of (8) with respect to ζ equal to zero and solving for ζ and a similar procedure then followed for ξ. Unfortunately, the form of the Hessian (7) means that it is difficult to obtain simple expressions for the required derivatives of $\log |\boldsymbol{A}|$.

D. The "Trick"

The difficulty in obtaining derivatives of $\log |\boldsymbol{A}|$ is due to the non-spherical nature of the prior distribution (5). We therefore transform the dual parameter space, so that the model is re-parameterised in order for the prior to become spherical, i.e. such that

$$\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} = \boldsymbol{\beta}^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_\ell\}$ is the vector of transformed parameters. The regularisation term is then equivalent to simple weight-decay. Let $\boldsymbol{G}$ represent the upper triangular Cholesky factor [22] of a symmetric positive-definite matrix $\boldsymbol{K}$, such that $\boldsymbol{K} = \boldsymbol{G}^T \boldsymbol{G}$. By inspection, the desired parameterisation is then given by

$$\boldsymbol{\beta} = \boldsymbol{G} \boldsymbol{\alpha} \quad \Longrightarrow \quad \boldsymbol{\alpha} = \boldsymbol{G}^{-1} \boldsymbol{\beta}.$$

The Bayesian analysis can then proceed using the evidence framework developed by [2–4], without further modification. The Hessian of S with respect to $\boldsymbol{\beta}$ can then be written as

$$\boldsymbol{A} = \boldsymbol{H} + \zeta \boldsymbol{I},$$

where $\boldsymbol{H} = \xi \nabla \nabla E_{\mathcal{D}}$ is the Hessian of $\xi E_{\mathcal{D}}$ with respect to $\boldsymbol{\beta}$. If the eigenvalues of $\boldsymbol{H}$ are $\lambda_1, \lambda_2, \ldots, \lambda_W$, then the eigenvalues of $\boldsymbol{A}$ are $(\lambda_1 + \zeta), (\lambda_2 + \zeta), \ldots, (\lambda_W + \zeta)$. The derivative of $\log |\boldsymbol{A}|$ with respect to ζ (assuming that the eigenvalues of $\boldsymbol{H}$ are independent of ζ) is then given by

$$\frac{d}{d\zeta} \log |\boldsymbol{A}| = \frac{d}{d\zeta} \log \left\{ \prod_{i=1}^{W} (\lambda_i + \zeta) \right\} = \sum_{i=1}^{W} \frac{1}{\lambda_i + \zeta}.$$

Setting the derivative of the log-evidence with respect to ζ to zero, we have that

$$2\zeta E_{\mathcal{W}}^{\rm MP} = W - \sum_{i=1}^{W} \frac{\zeta}{\lambda_i + \zeta} = \sum_{i=1}^{W} \frac{\lambda_i}{\lambda_i + \zeta} = \gamma,$$

where γ is the number of well determined parameters in the model. This leads to a simple update formula for the regularisation parameter:

$$\zeta^{\rm new} = \frac{\gamma}{2E_{\mathcal{W}}^{\rm MP}}. \quad (9)$$

Similarly at the maximum of the log-evidence (8) with respect to ξ,

$$2\xi E_{\mathcal{D}}^{\rm MP} = \ell - \sum_{i=1}^{W} \frac{\lambda_i}{\lambda_i + \zeta} = \ell - \gamma$$

giving the familiar update formula for the hyper-parameter ξ,

$$\xi^{\rm new} = \frac{\ell - \gamma}{2E_{\mathcal{D}}^{\rm MP}}. \quad (10)$$

The training procedure then alternates between updates of the primary model parameters, $\boldsymbol{\alpha}$, via (6) and updates of the hyper-parameters, ζ and ξ, according to equations (9-10). In practise, rather than using the transformed parameters only in computing the number of well-determined parameters, the entire training procedure is most easily conducted in the transformed parameters, $\boldsymbol{\beta}$, and the original model parameters, $\boldsymbol{\alpha}$, reclaimed afterwards. The values of kernel parameters can then be optimised by maximising the log-evidence for model, $\mathcal{H}_i$,

$$\begin{aligned} \log p(\mathcal{D}|\mathcal{H}_i) &= \zeta_{\rm MP} E_{\mathcal{W}}^{\rm MP} - \xi_{\rm MP} E_{\mathcal{D}}^{\rm MP} - \frac{1}{2} \log |\boldsymbol{A}| \\ &+ \frac{W}{2} \log \zeta_{\rm MP} + \frac{\ell}{2} \log \xi_{\rm MP} \\ &+ \frac{1}{2} \log \left\{ \frac{2}{\gamma} \right\} + \frac{1}{2} \log \left\{ \frac{2}{\ell - \gamma} \right\}, \end{aligned} \quad (11)$$

where in this case $\mathcal{H}_i$ specifies the kernel function (see [2] or [23] for further details).

E. Inducing Sparsity

The Gram matrix, $\boldsymbol{K}$, for a radial basis function kernel is at least in principle positive definite and of full rank, assuming that $\boldsymbol{x}_i \neq \boldsymbol{x}_j, \ \forall \ i, j \in \{1, 2, \ldots, \ell\}$ [24]; however it is possible for $\boldsymbol{K}$ to be *numerically* rank-deficient in which case the Cholesky factor $\boldsymbol{G}$ becomes ill-conditioned. We therefore use the incomplete Cholesky factorisation with symmetric pivoting, due to [1], to construct the Cholesky factor $\tilde{\boldsymbol{G}}$,

of $\tilde{\boldsymbol{K}}$, a numerically full-rank symmetric sub-matrix of $\boldsymbol{K}$. Again, without loss of generality, we assume that only the first W columns of $\boldsymbol{K}$ can contribute to forming $\tilde{\boldsymbol{G}}$; the remaining columns of $\boldsymbol{K}$ are then linearly dependent, or close to being linearly dependent, on columns $1, 2, \ldots, W$, and can be safely deleted prior to training without significantly affecting model performance (c.f. [25]). The Cholesky factor required to implement the re-parameterisation described in section II-D is then provided as a bye-product of the process used to identify redundant terms in the kernel expansion.

F. Alternative Approaches

The matrix $\boldsymbol{G}$ defining the required transformation is essentially the square-root of $\boldsymbol{K}$ and is not generally unique. Alternatively, we could use the *principal* square root of $\boldsymbol{K}$ [26]. Note both methods are commonly used to obtain a sample from an arbitrary multi-variate normal distribution from a sample drawn from a normal distribution with a unit covariance matrix. We adopt the incomplete Cholesky factorisation method [1] for four reasons:

- Both the principal square root has a computational complexity of $\mathcal{O}(\ell^3)$ operations. The incomplete Cholesky factorisation terminates when the linearly independent columns of $\boldsymbol{K}$ have been exhausted, as a result the computational complexity is only $\mathcal{O}(\ell W^2)$ operations, where W is the number of columns included in the Cholesky factor.
- The incomplete Cholesky factorisation provides a simple method of introducing sparsity.
- The Cholesky factor is derived from a *numerically* full-rank sub-matrix of $\boldsymbol{K}$ and so is highly likely to be stable.
- As the Cholesky factor is triangular, its inverse can be computed efficiently.

G. Sparse Bayesian Heteroscedastic Kernel Ridge Regression

Suppose we are given a dataset where the targets, t_i, are assumed to be realisations of some underlying function that have been corrupted by an independent and identically distributed[2] (i.i.d.) sample drawn from a Gaussian noise process with a mean of zero and input dependent variance,

$$t_i = \mu(\boldsymbol{x}_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma(\boldsymbol{x}_i)).$$

The conditional probability density of target t_i, given input vector $\boldsymbol{x}_i$ is given by

$$p(t_i|\boldsymbol{x}_i) = \frac{1}{\sqrt{2\pi}\sigma(\boldsymbol{x}_i)} \exp\left\{-\frac{[t_i - \mu(\boldsymbol{x}_i)]^2}{2\sigma^2(\boldsymbol{x}_i)}\right\}. \tag{12}$$

The negative log-likelihood (omitting constant terms) can then be written as

$$E_{\mathcal{D}} = \sum_{i=1}^{\ell} \left\{\log \sigma(\boldsymbol{x}_i) + \frac{[t_i - \mu(\boldsymbol{x}_i)]^2}{2\sigma^2(\boldsymbol{x}_i)}\right\}. \tag{13}$$

[2]By identically distributed we mean that the *conditional* distribution is identical for all samples, although the variance of the noise process is different for samples collected from different regions of $\mathcal{X}$

To model the data, we must jointly estimate the functions $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ using a kernel model [27]. The conditional mean is estimated by a linear model, $\mu(\boldsymbol{x}) = \boldsymbol{w}^\mu \cdot \boldsymbol{\phi}^\mu(\boldsymbol{x})$, constructed in a fixed feature space, $\mathcal{F}^\mu$ ($\boldsymbol{\phi}^\mu : \mathcal{X} \rightarrow \mathcal{F}^\mu$). Space $\mathcal{F}^\mu$ is induced by a positive definite kernel, $\mathcal{K}^\mu : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, defining the inner product $\mathcal{K}^\mu(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}^\mu(\boldsymbol{x}) \cdot \boldsymbol{\phi}^\mu(\boldsymbol{x}')$. The superscript μ is used to denote entities used to model the conditional mean $\mu(\boldsymbol{x})$. The standard deviation is a strictly positive quantity and so the *logarithm* of the standard deviation is estimated by a second linear model, $\log \sigma(\boldsymbol{x}_i) = \boldsymbol{w}^\sigma \cdot \boldsymbol{\phi}^\sigma(\boldsymbol{x})$, similarly constructed in a feature space $\mathcal{F}^\sigma$ defined by Mercer kernel $\mathcal{K}^\sigma$. A superscript σ is used to identify entities used to model the standard deviation, $\sigma(\boldsymbol{x})$. The parameters of the model ($\boldsymbol{w}^\mu$, and $\boldsymbol{w}^\sigma$) are determined by minimising the objective function

$$\begin{aligned} S(\boldsymbol{w}^\mu, \boldsymbol{w}^\sigma) &= \sum_{i=1}^{\ell} \left\{\log \sigma(\boldsymbol{x}_i) + \frac{[t_i - \mu(\boldsymbol{x}_i)]^2}{2\sigma^2(\boldsymbol{x}_i)}\right\} \\ &+ \frac{1}{2}\zeta^\mu \|\boldsymbol{w}^\mu\|^2 + \frac{1}{2}\zeta^\sigma \|\boldsymbol{w}^\sigma\|^2. \end{aligned} \tag{14}$$

Again the representer theorem means that these models can be represented by kernel expansions. In this case, we approximate the full kernel model using sparse expansions, such that

$$\mu(\boldsymbol{x}) = \sum_{i=1}^{W^\mu} \alpha_i^\mu \mathcal{K}^\mu(\boldsymbol{x}, \boldsymbol{x}_i),$$

and

$$\log \sigma(\boldsymbol{x}) = \sum_{i=1}^{W^\sigma} \alpha_i^\sigma \mathcal{K}^\sigma(\boldsymbol{x}, \boldsymbol{x}_i).$$

Again, the model is re-parameterised to facilitate a Bayesian analysis under the evidence framework, such that

$$\boldsymbol{\beta}_\mu = \boldsymbol{G}_\mu \boldsymbol{\alpha}^\mu \quad \text{and} \quad \boldsymbol{\beta}_\sigma = \boldsymbol{G}_\sigma \boldsymbol{\alpha}^\sigma,$$

where $\boldsymbol{G}_\mu^T \boldsymbol{G}_\mu = \boldsymbol{K} = \{\mathcal{K}^\mu(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^{W^\mu}$ and $\boldsymbol{G}_\sigma^T \boldsymbol{G}_\sigma = \boldsymbol{K}^\sigma = \{\mathcal{K}^\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^{W^\sigma}$. In this case the posterior distribution over the model parameters, $\boldsymbol{w}^\mu$ and $\boldsymbol{w}^\sigma$, is only approximated by a Gaussian centred on the most probable values. Under this approximation, the log-evidence for the hyper-parameters, ζ^μ and ζ^σ, is given by

$$\begin{aligned} \log p(\mathcal{D}|\zeta^\mu, \zeta^\sigma) &= \frac{W^\mu}{2}\log \zeta^\mu + \frac{W^\sigma}{2}\log \zeta^\mu \\ &- \frac{\ell}{2}\log(2\pi) - E_{\mathcal{D}}^{\mathrm{MP}} \\ &- \zeta^\mu E_{\mathcal{W}^\mu}^{\mathrm{MP}} - \frac{1}{2}\log|\boldsymbol{A}^\mu| \\ &- \zeta^\sigma E_{\mathcal{W}^\sigma}^{\mathrm{MP}} - \frac{1}{2}\log|\boldsymbol{A}^\sigma|, \end{aligned} \tag{15}$$

where $E_{\mathcal{W}^\mu} = \sum_{i=1}^{W^\mu} \beta_i^\mu$, $E_{\mathcal{W}^\sigma} = \sum_{i=1}^{W^\sigma} \beta_i^\sigma$ and $\boldsymbol{A}^\mu$ and $\boldsymbol{A}^\sigma$ represent the Hessian matrices of (14) with respect to $\boldsymbol{\beta}^\mu$ and $\boldsymbol{\beta}^\sigma$ respectively. Update formula for the regularisation parameters can then be obtained as before,

$$\zeta_{\mathrm{new}}^\mu = \frac{\gamma^\mu}{2E_{\mathcal{W}^\mu}}, \qquad \gamma^\mu = \sum_{i=1}^{W^\mu} \frac{\lambda_i^\mu}{\lambda_i^\mu + \zeta^\mu}$$

and

$$\zeta^{\sigma}_{\text{new}} = \frac{\gamma^{\sigma}}{2E_{\mathcal{W}^{\sigma}}}, \qquad \gamma^{\sigma} = \sum_{i=1}^{W^{\sigma}} \frac{\lambda^{\sigma}_i}{\lambda^{\sigma}_i + \zeta^{\sigma}},$$

where $\boldsymbol{\lambda}^{\mu}$ and $\boldsymbol{\lambda}^{\sigma}$ represent the eigenvalues of the Hessian matrices of $E_{\mathcal{D}}$ with respect to $\boldsymbol{\beta}^{\mu}$ and $\boldsymbol{\beta}^{\sigma}$ respectively. Good values for the kernel parameters can be obtained by maximising the log-evidence for the model $\mathcal{H}_i$, given by an expression analogous to (11).

III. Results

In this section, we compare Bayesian, cross-validation based and hybrid model selection strategies for conventional and heteroscedastic kernel ridge regression models of the Motorcycle benchmark datasets. The Bayesian strategy takes an evidence-based approach to optimisation of all regularisation, inverse noise variance and kernel parameters. The cross-validation strategy chooses good values for all of these parameters by minimising a 10-fold cross-validation estimate of the negative log-likelihood. The hybrid strategy optimises regularisation and inverse noise variance parameters via the Bayesian approach and the kernel parameters via 10-fold cross-validation. The Nelder-Mead simplex algorithm [28] was used for all function minimisation problems. The Motorcycle benchmark consists of a sequence of accelerometer readings through time following a simulated motorcycle crash performed during experiments to determine the efficacy of crash helmets [29]. Figure 1 shows conventional and heteroscedastic kernel ridge regression models resulting from Bayesian and cross-validation based model selection strategies; both strategies are seen to produce sensible models of the data with little indication of over-fitting.

Table I shows leave-one-out cross-validation estimates of the sum-of-squares error and negative log-likelihood given by (13) for conventional and heteroscedastic kernel ridge regression models for each model selection strategy. The differences between model selection strategies in terms of generalisation seem relatively slight, none of the model selection strategies clearly dominate any of the statistics considered.

TABLE I

Leave-one-out cross-validation estimates of sum-of-squares error and negative log-likelihood for the Motorcycle benchmark dataset [29].

Model	Selection	SSE	$E_{\mathcal{D}}$
KRR	Cross-validation	70851	486.41
KRR	Bayesian	72150	487.06
KRR	Hybrid	70929	485.95
HKRR	Cross-validation	72100	439.70
HKRR	Bayesian	71269	462.39
HKRR	Hybrid	74513	453.58

Table II shows the number of kernel functions used to model the conditional mean and conditional variance of the target distribution and also the time taken for model selection. The advantages of the Bayesian model selection strategy become apparent; the Bayesian strategy is not only significantly faster than cross-validation and hybrid strategies, but also results in the most compact models with fewest parameters.

TABLE II

Number of basis functions used to model the conditional mean, W^{μ} and conditional variance, W^{σ}, and model selection time for the motorcycle benchmark dataset [29].

Model	Selection	W^{μ}	W^{σ}	Time
KRR	Cross-validation	27	—	26.24 s
KRR	Bayesian	16	—	1.09 s
KRR	Hybrid	20	—	5.61 s
HKRR	Cross-validation	25	32	176.92 s
HKRR	Bayesian	18	6	12.58 s
HKRR	Hybrid	15	14	154.11 s

IV. Conclusion

In this paper we have introduced a simple mathematical device simplifying the construction of Bayesian treatments of a family of sparse kernel learning methods. The use of this technique was demonstrated by a Bayesian analysis of sparse conventional and heteroscedastic kernel ridge regression algorithms. The limited simulation results presented indicate that Bayesian model selection strategies are competitive with cross-validation based approaches, with reduced computational expense. It is likely that this method will be more widely applicable, for non-linear regression with non-Gaussian noise models, for example in parametric survival analysis [30] or in modelling rainfall [31].

Acknowledgements

The authors would like to thank Andy Hanna and Katya Scheinberg for their helpful advice and assistance in implementing the incomplete Cholesky factorisation algorithm. This work was supported by a grant from the Biotechnology and Biological Sciences Research Council (grant number 83/D17534).

References

[1] S. Fine and K. Scheinberg, "Efficient SVM training using low-rank kernel representations," *Journal of Machine Learning Research*, vol. 2, pp. 243–264, Dec. 2001.

[2] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[3] D. J. C. MacKay, "The evidence framework applied to classification networks," *Neural Computation*, vol. 4, no. 5, pp. 720–736, 1992.

[4] D. J. C. MacKay, "A practical Bayesian framework for backprop networks," *Neural Computation*, vol. 4, pp. 448–472, 1992.

[5] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *Proc., 15th Int. Conf. on Machine Learning*, Madison, WI, July 24–27 1998, pp. 515–521.

[6] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[7] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[8] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*. 1999, pp. 41–48, IEEE Press.

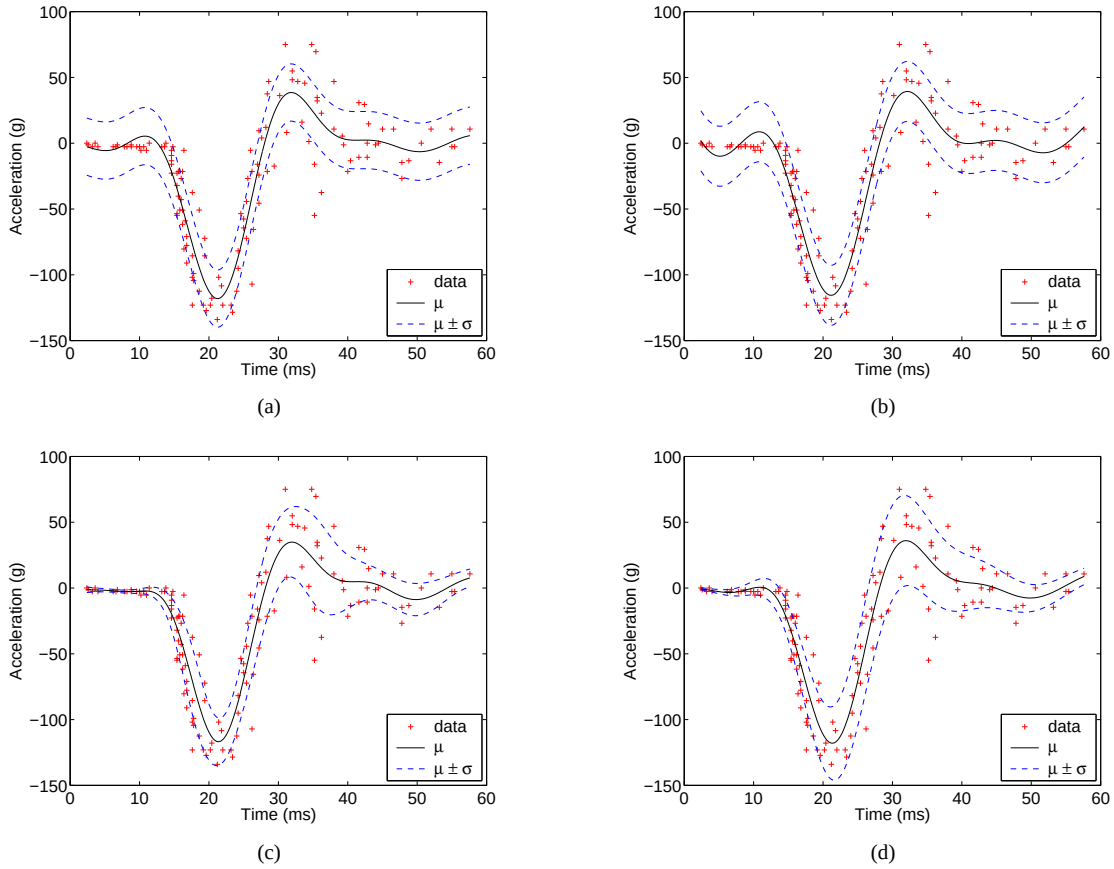

Fig. 1. (a) Kernel ridge regression model of the Motorcycle benchmark dataset [29] with model selection based on 10-fold cross-validation, (b) kernel ridge regression with Bayesian model selection, (c) heteroscedastic kernel ridge regression with 10-fold cross-validation based model selection and (d) heteroscedastic kernel ridge regression with Bayesian model selection.

[9] I. T. Jolliffe, *Principal component analysis*, Springer, second edition, 2002.

[10] B. Schölkopf, A. J. Smola, and Müller K.-R., "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[11] B. E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth ACM Workshop on Computational Learning Theory*, July 1992, pp. 144–152.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sept. 1995.

[13] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society, B*, vol. 36, no. 1, pp. 111–147, 1974.

[14] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 131–159, 2002.

[15] J. T.-Y. Kwok, "The evidence framework applied to support vector machines," *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp. 1162–1173, Sept. 2000.

[16] T. Van Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle, "Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis," *Neural Computation*, vol. 14, no. 5, pp. 1115–1147, 2002.

[17] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*, John Wiley, New York, 1977.

[18] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[19] G. S. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *J. Math. Anal. Applic.*, vol. 33, pp. 82–95, 1971.

[20] W. L. Buntine and A. S. Weigend, "Bayesian back-propagation," *Complex Systems*, vol. 5, pp. 603–643, 1991.

[21] D. J. C. MacKay, "Hyperparameters : optimise or integrate out?," in *Maximum Entropy and Bayesian Methods*. Kluwer, 1994.

[22] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, third edition edition, 1996.

[23] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[24] C. A. Micchelli, "Interpolation of scattered data: Distance matrices and conditionally positive definite functions," *Constructive Approximation*, vol. 2, pp. 11–22, 1986.

[25] G. Baudat and F. Anouar, "Kernel-based methods and function approximation," in *Proc. IJCNN*, Washington, DC, July 2001, pp. 1244–1249.

[26] N. J. Higham, "Computing real square roots of a real matrix," *Linear algebra and applications*, vol. 88/89, pp. 405–430, 1987.

[27] G. C. Cawley, N. L. C. Talbot, R. J. Foxall, S. R. Dorling, and D. P. Mandic, "Heteroscedastic kernel ridge regression," *Neurocomputing*, vol. 57, pp. 105–124, 2004.

[28] J. A. Nelder and R. Mead, "A simplex method for function minimisation," *Computer Journal*, vol. 7, pp. 308–313, 1965.

[29] B. W. Silverman, "Some aspects of the spline smoothing approach to non-parametric regression curve fitting," *J. Royal Statistical Society, B*, vol. 47, no. 1, pp. 1–52, 1985.

[30] D. R. Cox and D. Oakes, *Analysis of Survival Data*, Chapman and Hall, London, 1984.

[31] P. M. Williams, "Modelling seasonality and trends in daily rainfall data," in *Advances in Neural Information Processing Systems*, vol. 10, pp. 985–991. MIT Press, 1998.